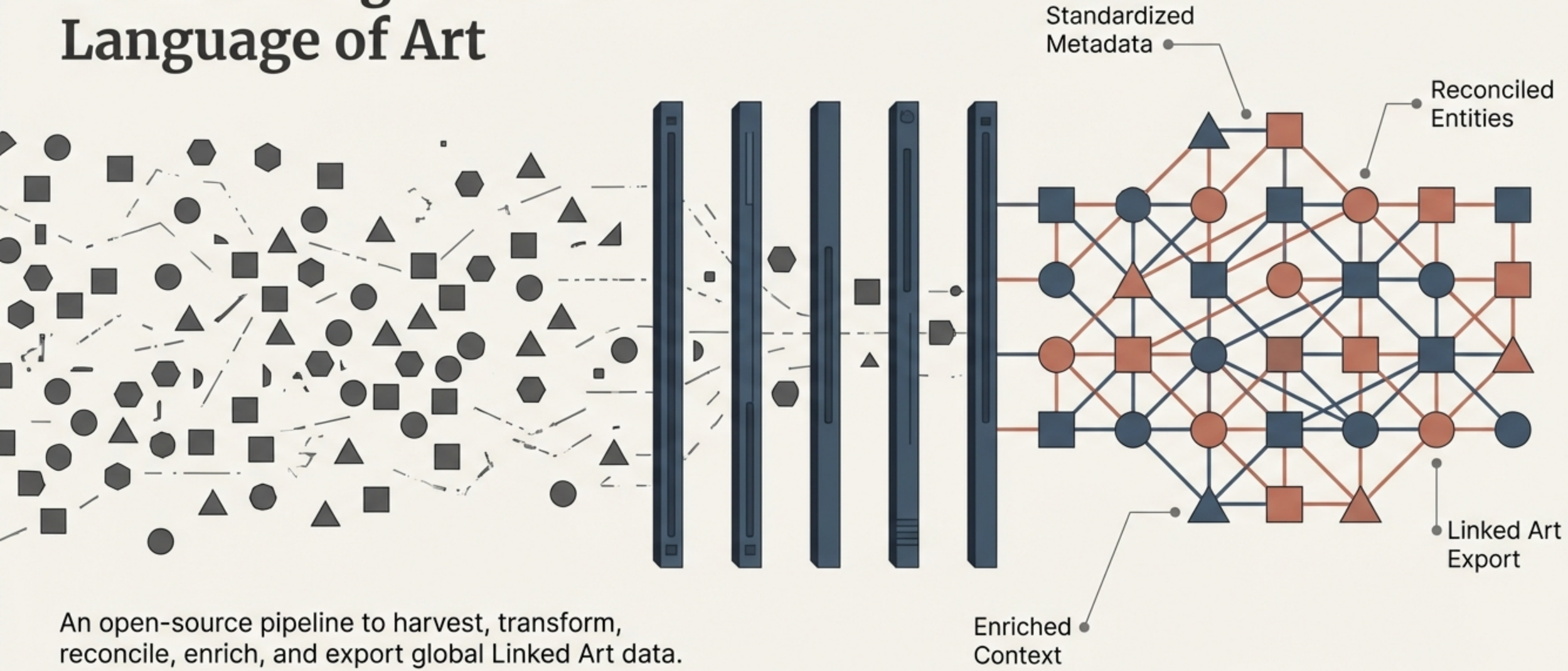
Project LUX Data Pipeline

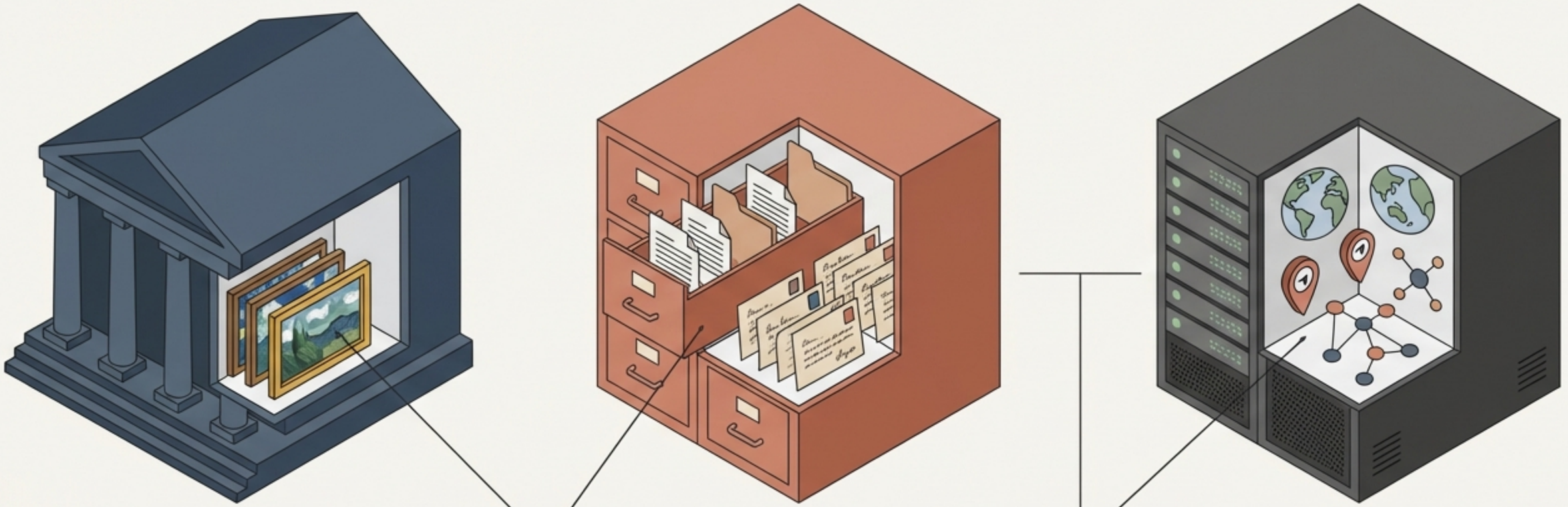
# Architecting the Universal Language of Art




An open-source pipeline to harvest, transform, reconcile, enrich, and export global Linked Art data.

NotebookLM

# Cultural heritage data is currently trapped in isolated, incompatible silos



**The Van Gogh Problem:** Paintings are tracked in one museum's system, letters in another archive, and historical context scattered across dozens of external databases.

The result is a nightmare of information overload—conflicting formats, different languages, and tedious manual cross-referencing for researchers.

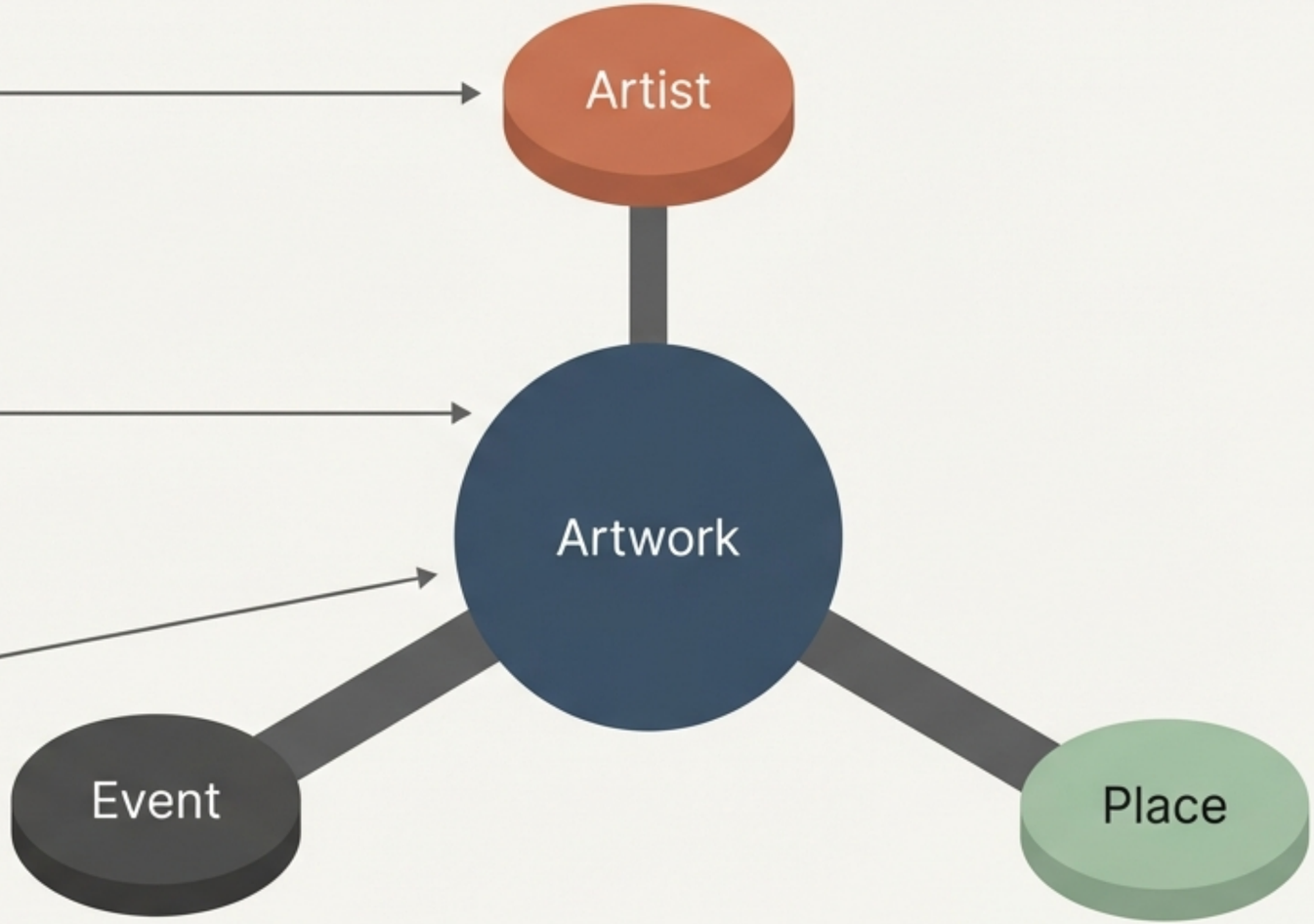
NotebookLM

# Linked Art provides a universal standard to connect global cultural heritage

**Permanent Web Addresses:** Every piece of data receives a unique, permanent identifier.

**Meaningful Lines:** Machines can read the exact relationships between an artwork, an artist, and a historical event.

**Universal Language:** Replaces quirky, proprietary museum formats with a single structural model.




NotebookLM

# The LUX pipeline operates as an automated factory for data integration

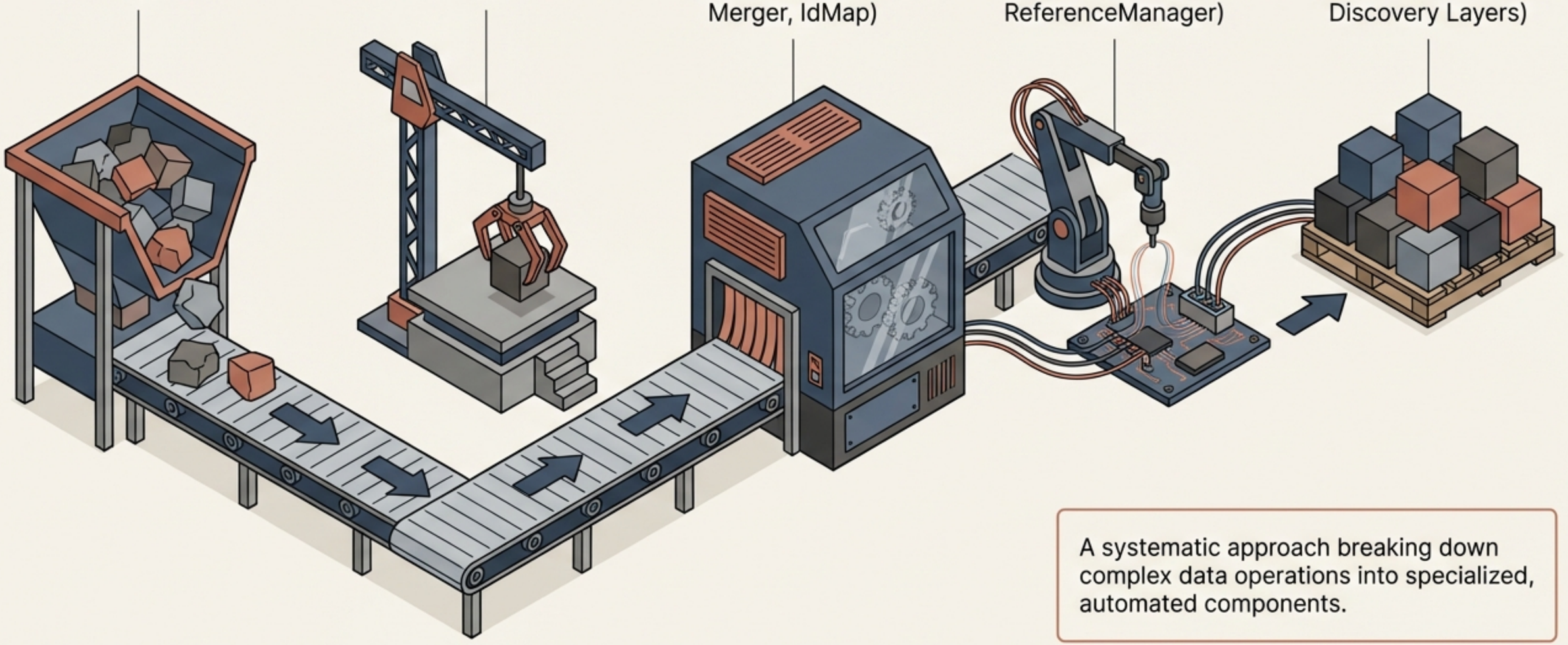
**Raw Materials**
(External Sources)

**Loading Dock**
(Harvester & Caches)

**Core Processing**
(Collector, Reconciler, Merger, IdMap)

**Internal Wiring**
(Reidentifier, ReferenceManager)

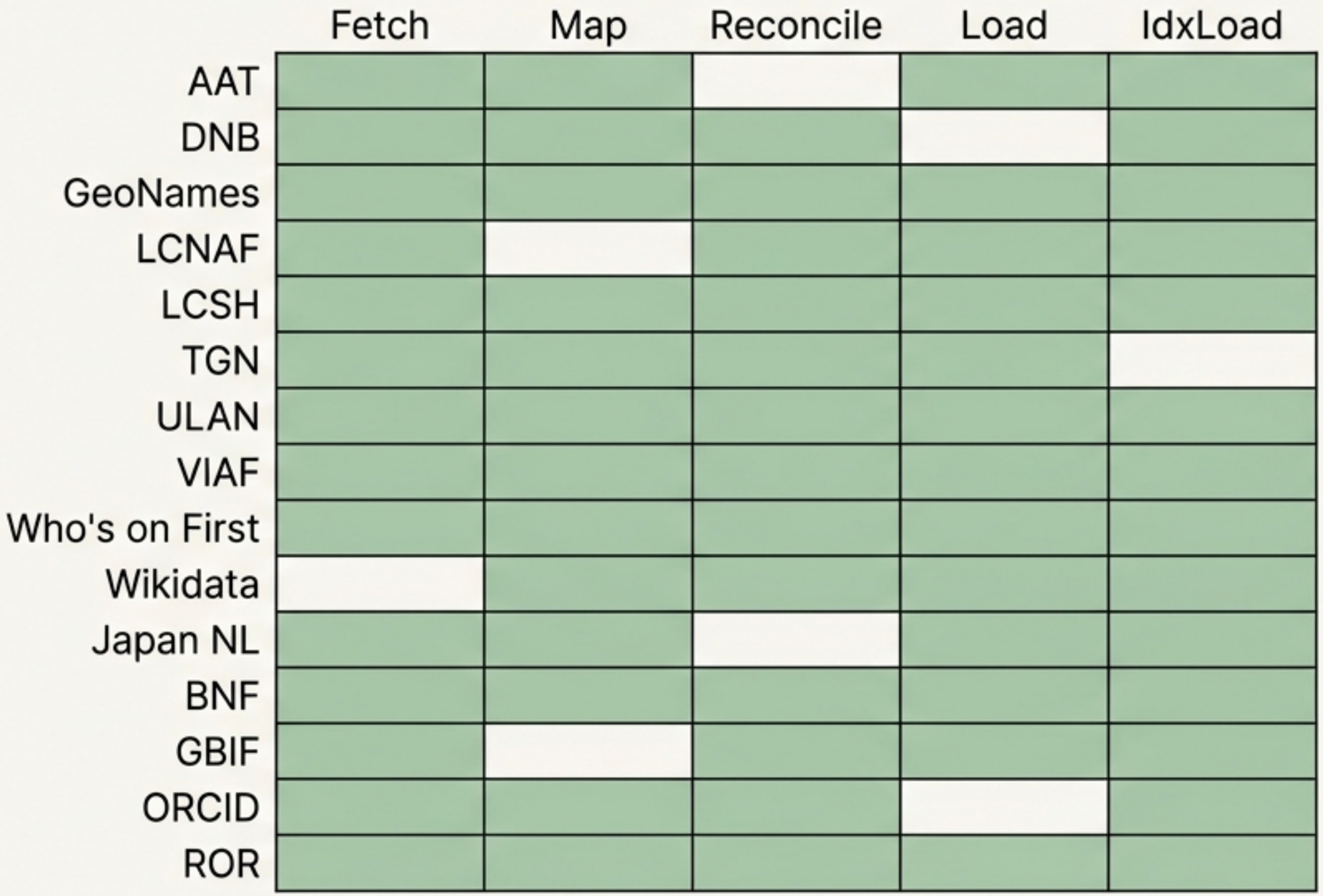
**Export**
(MarkLogic Storage, Discovery Layers)



A systematic approach breaking down complex data operations into specialized, automated components.

NotebookLM

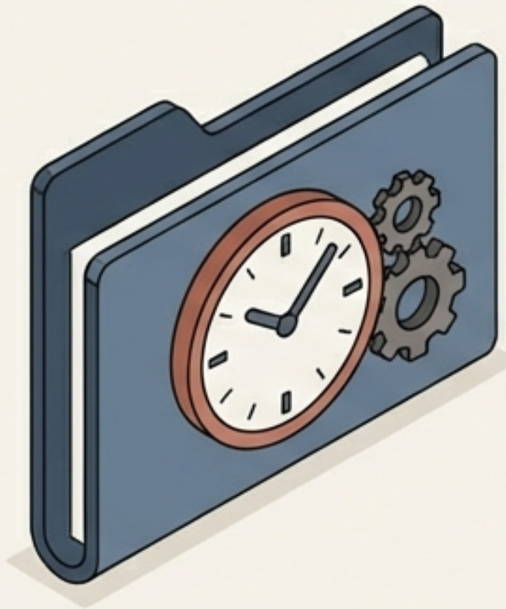
# The pipeline integrates an unprecedented scale of global authority sources




Balancing Effort vs. Reward: Sources like FAST are deliberately deferred because parsing their legacy MARC/XML formats requires resources better spent on broader integrations.

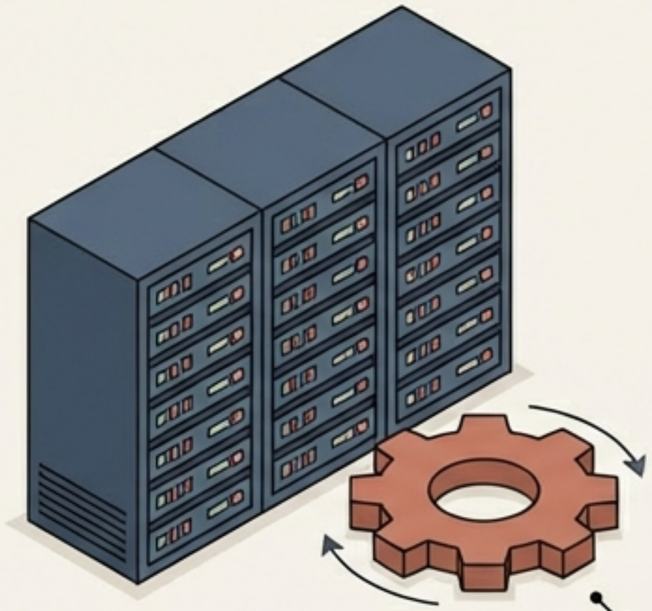
NotebookLM

# Pragmatic engineering handles massive, continuous data ingestion



## Version Control for Data Dumps

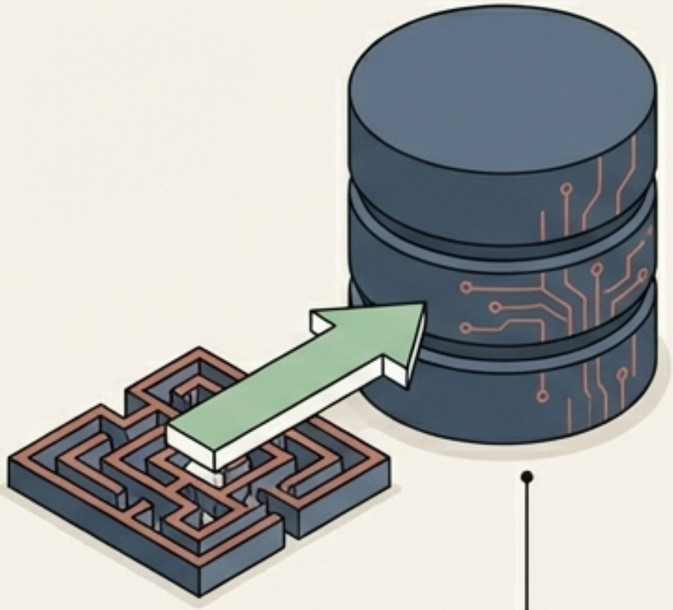
The pipeline looks up `dumpFilePath` and archives old local data with date stamps (e.g., `latest-2022-07`) before fetching new files via `wget`.



## Backgrounding the Behemoths

```
nohup wget --quiet <url> &
```

Used for enormous files like Wikidata, allowing the pipeline to continue working without waiting hours for downloads.

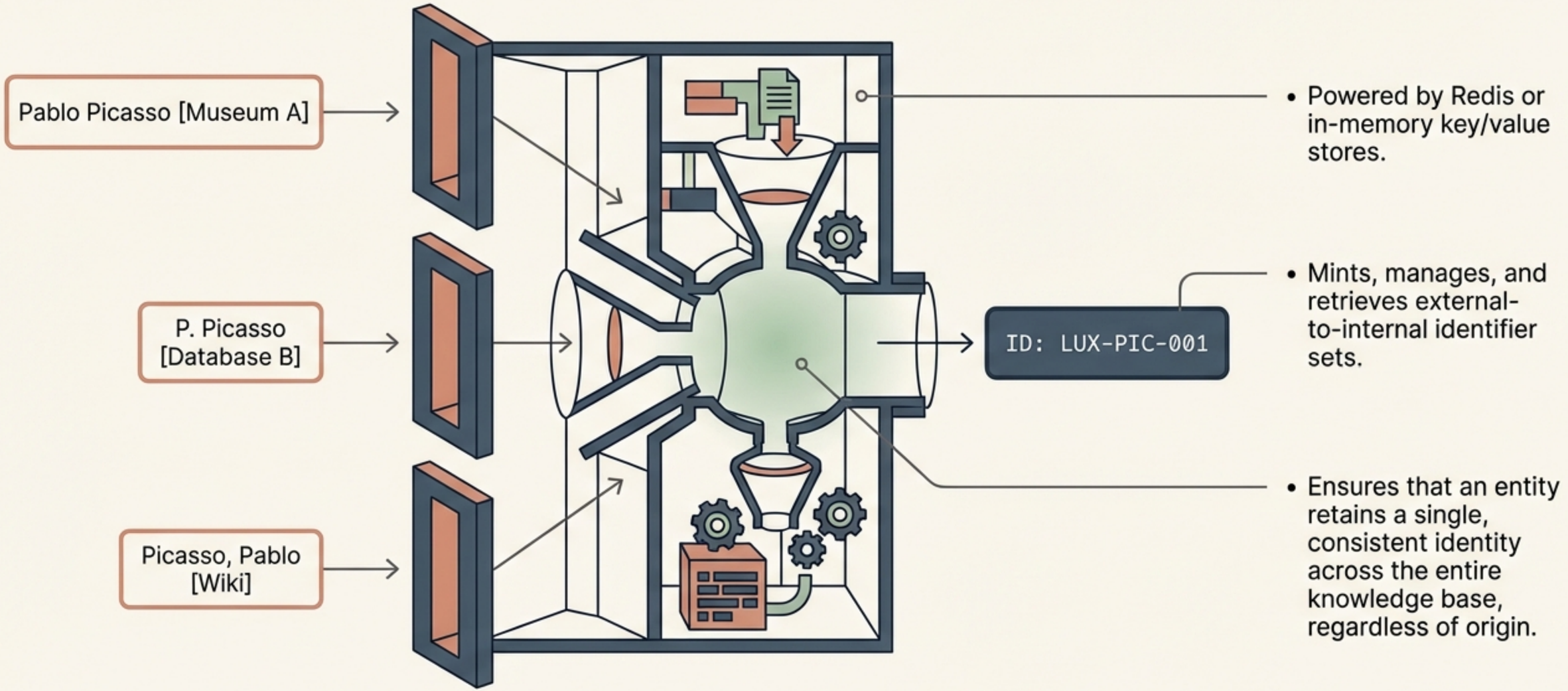


## Bypassing the Build

Rather than unpacking the massive 33GB Who's on First SQLite database, the pipeline efficiently treats the `.sqlite` file itself directly as a cache.

NotebookLM

# The IdMap acts as a master directory to ensure consistent identities

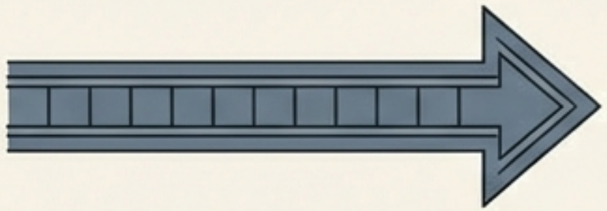

Pablo Picasso [Museum A]
P. Picasso [Database B]
Picasso, Pablo [Wiki]
ID: LUX-PIC-001
• Powered by Redis or in-memory key/value stores.
• Mints, manages, and retrieves external-to-internal identifier sets.
• Ensures that an entity retains a single, consistent identity across the entire knowledge base, regardless of origin.

NotebookLM

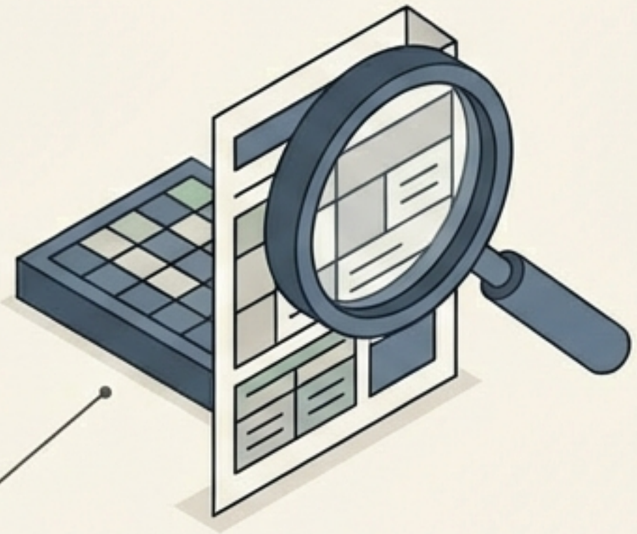
# The core assembly line systematically gathers and verifies every record

`process/Collector`

Acts as an aggressive gatherer, recursively collecting every related identifier for a given record to ensure no data fragment is left behind.



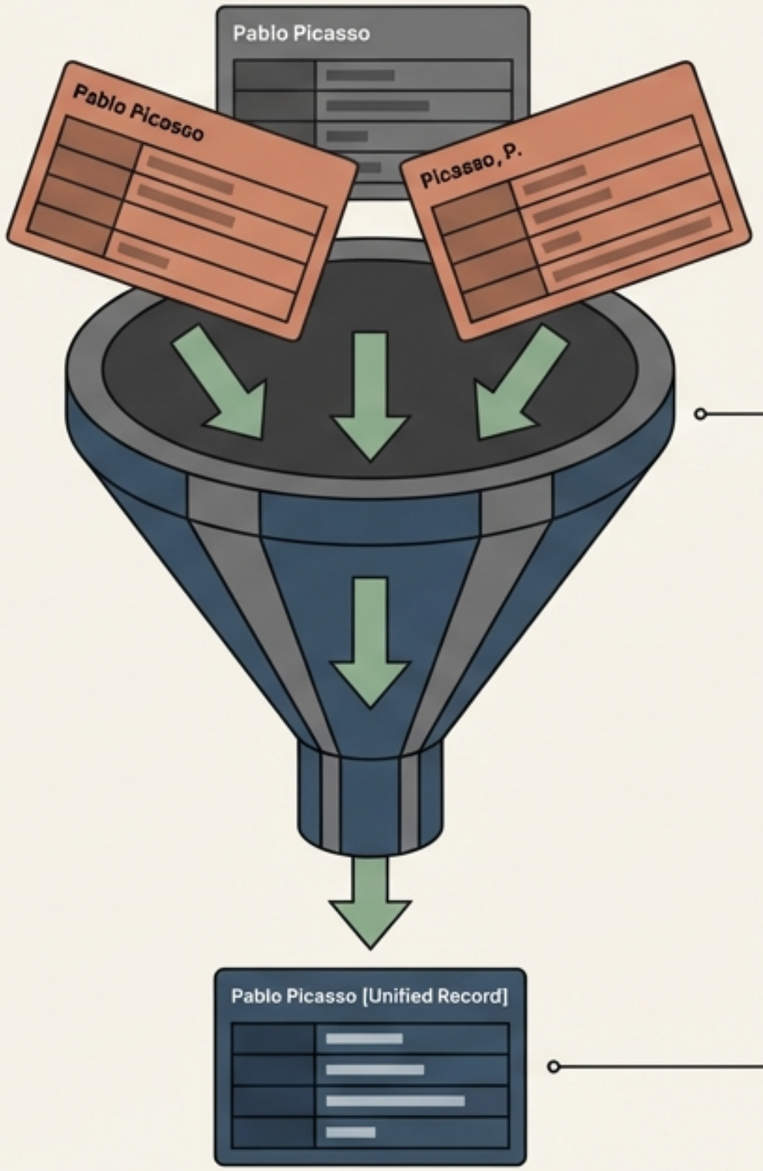
`sources/*/Reconciler`



Cross-references incoming entities against external authority files (like Getty ULAN or Wikidata) to verify if the entity is already documented, preventing redundant processing.

NotebookLM

# The Merger acts as a smart consolidator for conflicting or redundant data




## The Unsung Hero: process/Merger

- Intelligently combines disparate Linked Art records representing the exact same entity.
- Eliminates redundancy and resolves conflicting information.
- Saves researchers countless hours by presenting one definitive, clean entry rather than multiple overlapping fragments.

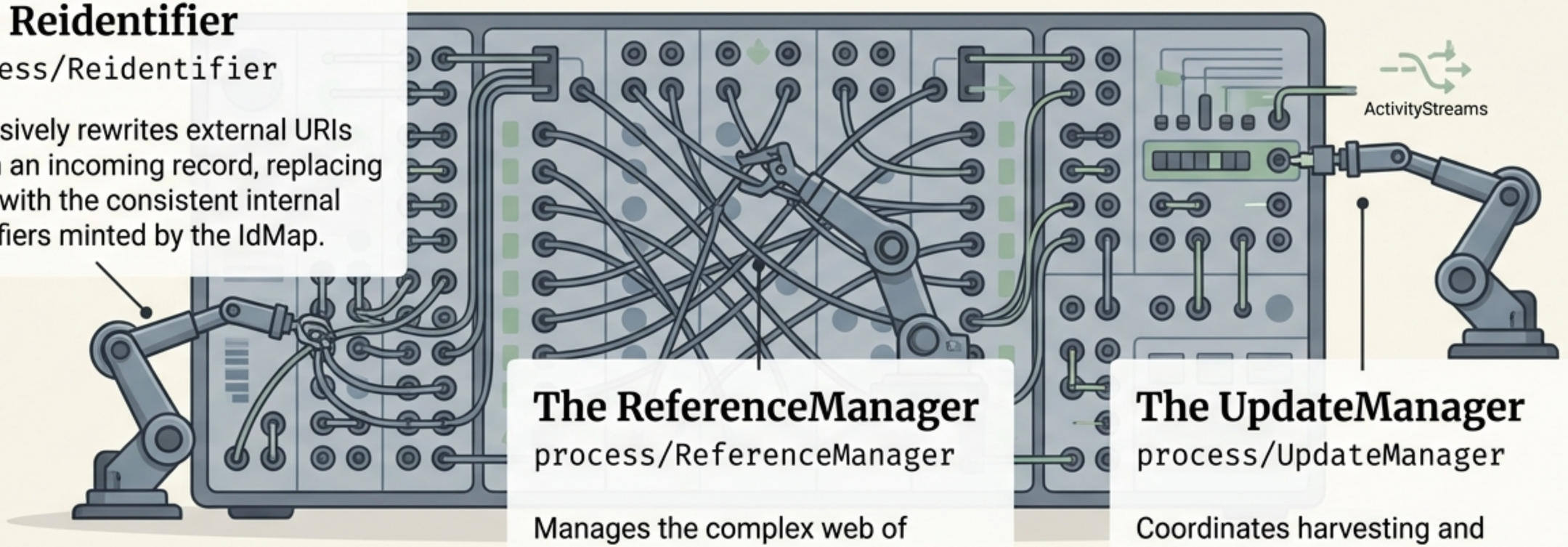
NotebookLM

# Internal wiring ensures every historical connection remains perfectly intact

## The Reidentifier

`process/Reidentifier`

Recursively rewrites external URIs within an incoming record, replacing them with the consistent internal identifiers minted by the IdMap.




## The ReferenceManager

`process/ReferenceManager`

Manages the complex web of inter-record connections, ensuring that when an entity is updated, all inbound and outbound links remain perfectly aligned.

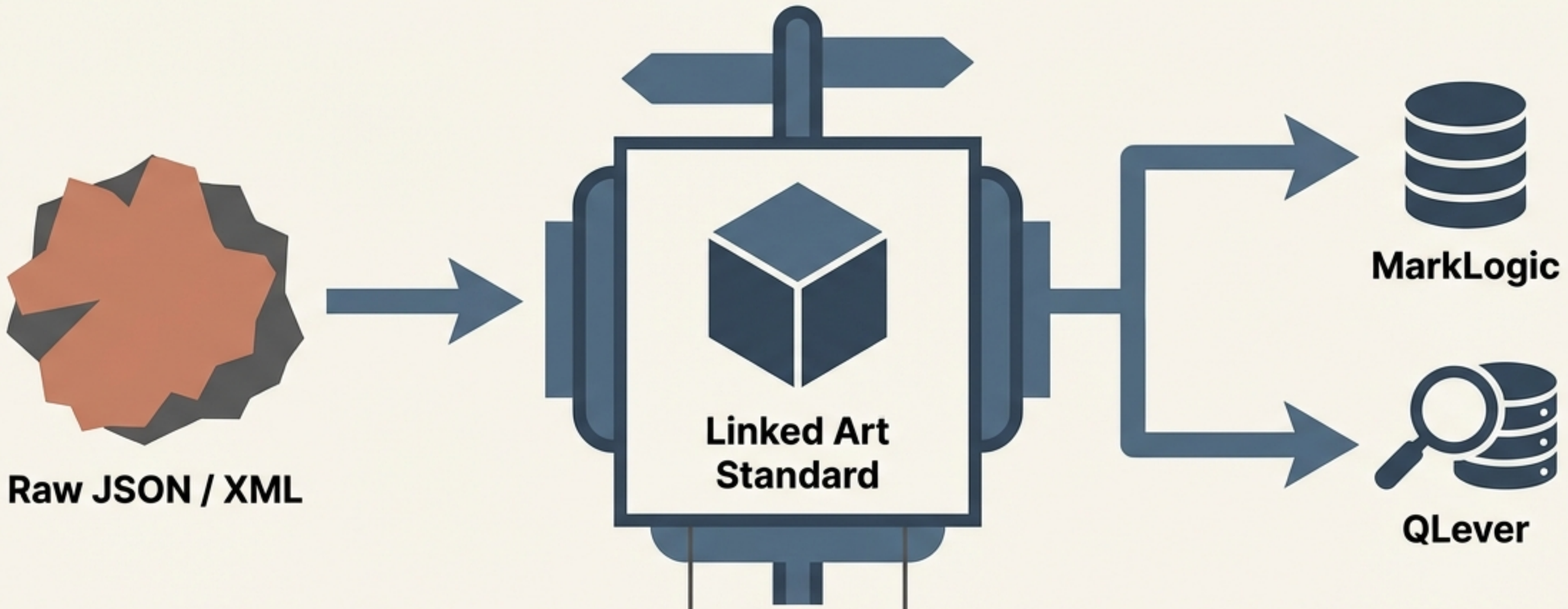
## The UpdateManager

`process/UpdateManager`

Coordinates harvesting and continuously keeps local caches fresh, guided by ActivityStreams (a standardized news feed for data changes).

NotebookLM

# The Mapper translates raw factory output into discovery-ready formats

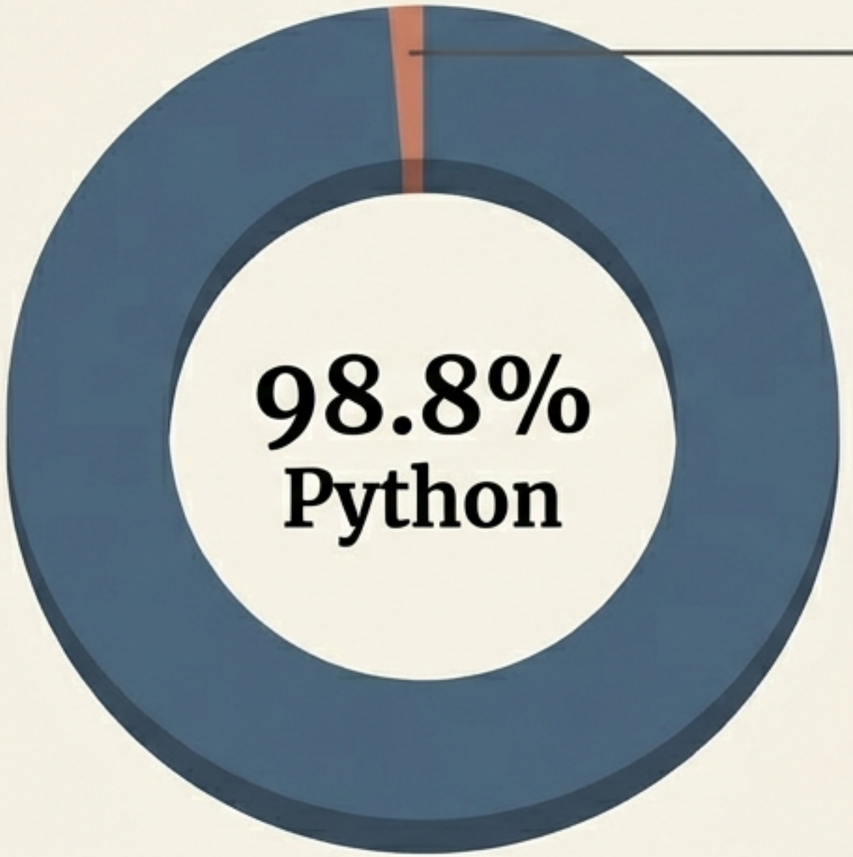



The pipeline's master translator. `sources/*/Mapper` first converts quirky museum source formats into standardized **Linked Art**.

It then performs a second critical translation, formatting that Linked Art data specifically for discovery layers and persistent storage systems like `MarkLogic` and `QLever`.

NotebookLM

# A modular Python codebase drives the entire pipeline architecture




Leveraging Python’s unparalleled flexibility for complex data manipulation.

### 1.2% Shell Scripting

Utilized for parallel processing and highly efficient system-level tasks.

- **Configuration-Driven:** Pipeline logic is dictated by JSON records stored in cache.
- **Smart Syncing:** Utilizes IIIF Change Discovery to efficiently synchronize enormous N-Triples datasets (like Getty AAT/TGN/ULAN) without rebuilding billions of rows from scratch.

NotebookLM

# Researchers can move instantly from manual data wrangling to deep analysis

## Answering Bigger Questions:

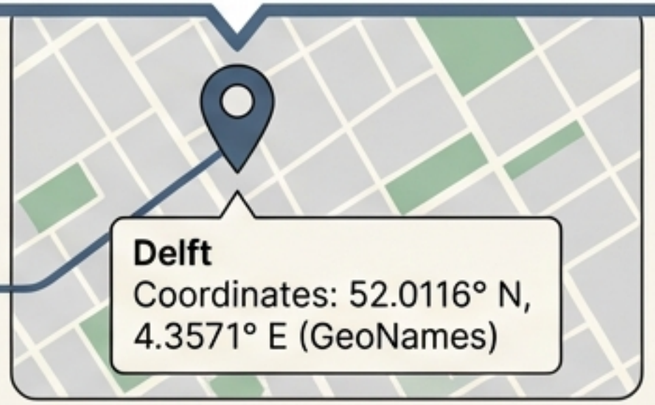
Instead of spending months cross-referencing names, historians can instantly analyze complex queries—such as tracing how political events in 17th-century Amsterdam influenced subject matter across all of Northern Europe.

**Johannes Vermeer**
(Rijksmuseum)

**Jan Vermeer**
(Genealogy Database)

**View of Delft**

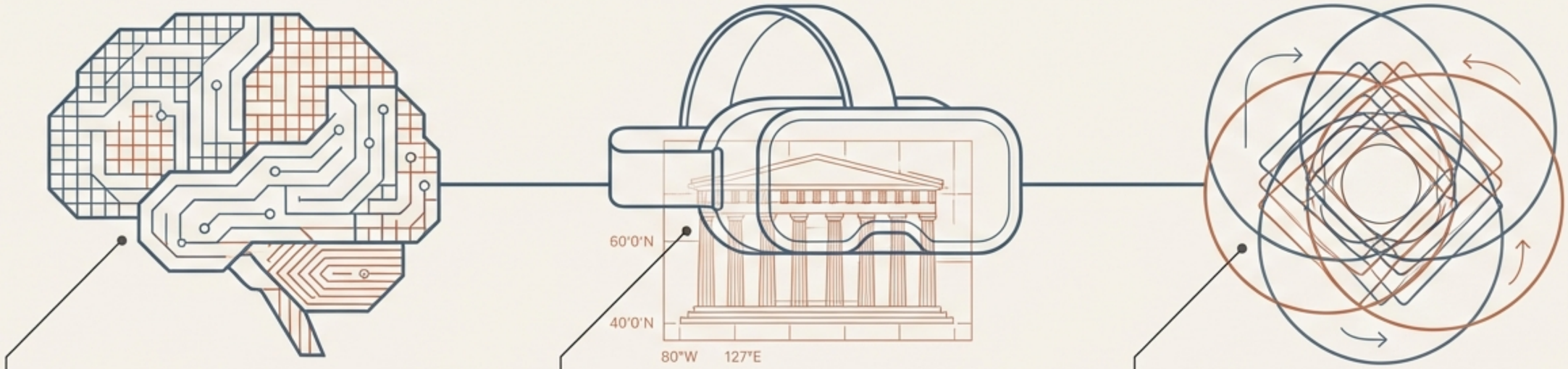
Instant Linked Data Access



**Delft**
Coordinates: 52.0116° N, 4.3571° E (GeoNames)

NotebookLM

# This structured data foundation enables next-generation applications




## AI & Machine Learning

Provides the incredibly clean, structured, and interconnected training data required for AI to accurately analyze complex artistic styles and historical trends.

## Immersive Reality (VR/AR)

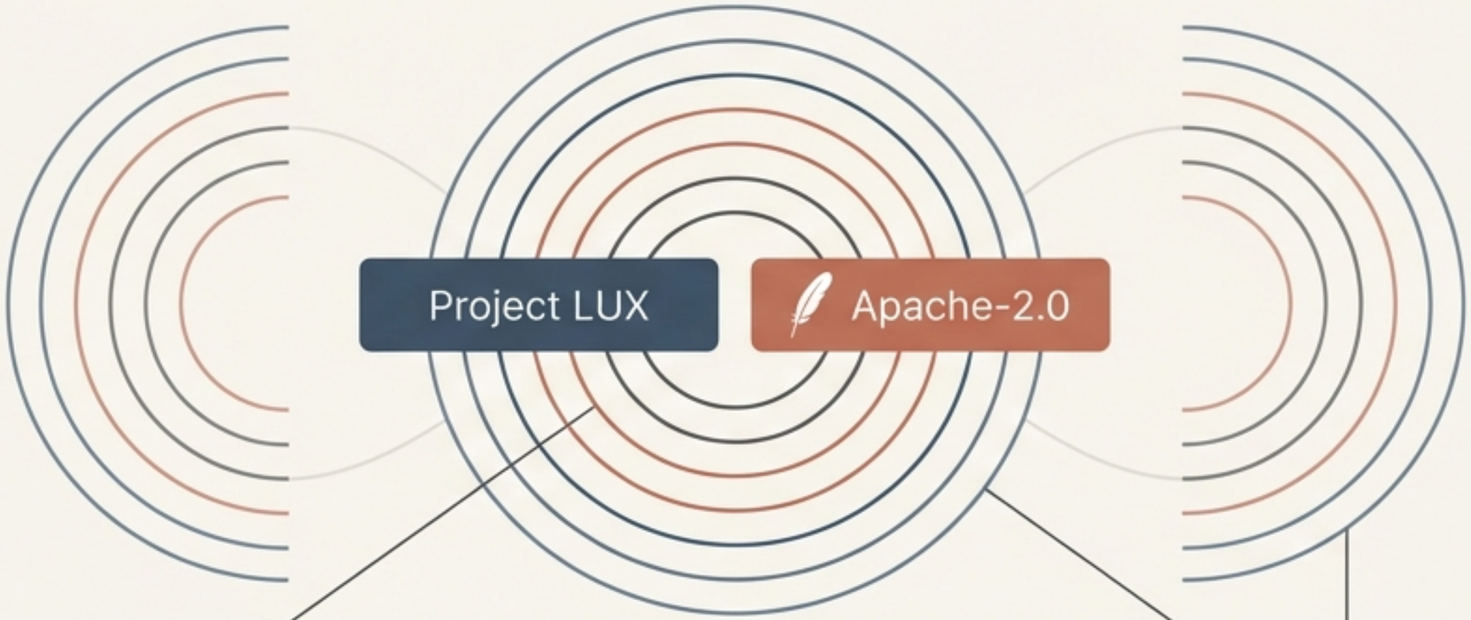
Powers virtual reality experiences that bring historical sites to life by pulling in perfectly reconciled, real-time geographical and architectural link data.

## Cross-Domain Adaptation

The modular architecture of automated harvesting, mapping, and reconciling can be directly adapted to untangle data deluges in scientific research or market analysis.

NotebookLM

# The LUX pipeline is a blueprint for solving complex information deluge


Project LUX
Apache-2.0
• An invisible backbone transforming chaotic data into coherent, usable knowledge.
• Available on GitHub as a fully modular, open-source asset to the digital humanities community.

*"By integrating the world's most robust authority sources, it does not just store data—it distills vast amounts of information into a navigable,* ***living knowledge base****."*

NotebookLM